https://doi.org/10.31272/jae.i150.1447
https://admics.uomustansiriyah.edu.iq
P-ISSN: 1813–6729 E-ISSN: 2707-1359

Fitting the Lindley Survival Model Using Maximum Likelihood and Moments Method for Analysing Leukemia Survival Data in Erbil_Iraq

Ibrahim A. Othman
Dept. of Statistics & Informatics, College of Administration and Economics, Salahaddin University, Erbil, Kurdistan, Iraq.
Email: ibrahim.othman@su.edu.krd, ORCID: https://orcid.org/0009-0000-2179-9024
Kurdistan I. Mawlood
Dept. of Statistics & Informatics, College of Administration and Economics, Salahaddin University, Erbil, Kurdistan, Iraq.
Email: kurdistan.mawlood@su.edu.krd, ORCID: https://orcid.org/0000-0002-1612-1996

Article Information

Article History:
Received: 23 / 06 / 2025
Revised: 01 / 11 / 2025
Accepted: 06 / 11 / 2025
Available Online: 01 / 12 / 2025
Pages no: 39 – 53

Keywords:
Survival Analysis, Lindley parametric survival model, Akaike Information Criterion (AIC), Leukemia.

Correspondence:
Researcher name:
Kurdistan I. Mawlood

Email:
kurdistan.mawlood@su.edu.krd

Abstract

The study uses the Lindley distribution to estimate parametric survival models using leukemia survival data in Erbil City, Iraq. We examine two estimation methods for modelling time-to-event data: the Maximum Likelihood estimate (MLE) and the Method of Moments (MoM). The MLE approach produces asymptotically efficient estimates, whereas using sample moments, the MoM provides a more straightforward, non-iterative approach. The goodness of fit tests, Akaike Information Criteria (AIC), Bayesian Information Criterion (BIC) and Mean Square Error (MSE) are used to assess model performance.

The results show that MLE outperforms MoM in terms of precision and resilience, especially with censored data. However, the findings demonstrate the relevance of the Lindley distribution in medical survival analysis and provide insight into leukemia survival patterns in Erbil City.

This study adds to the growing literature on parametric survival modelling and helps clinical decision-making for leukemia therapies. The data set of this study was obtained from Rizgari Hospital in Erbil city, the results obtained by utilising the statistical packages (MATLAB V. 23.2 and R Software V. 2025.05.1+513).

1. Introduction

The foundation of survival data analysis, an essential issue in health, particularly for leukemia patients, includes the fundamental concepts and basic methods for modelling survival data in the presence of censored observations.

Survival time is defined as the time from some starting point to some endpoint of the subject's life. Survival, of course, can be understood figuratively. In this case, survival refers to the individual being in a state that aligns with the baseline state. The situation won't change until an interesting thing happens. Failure is the crucial event that marks the conclusion of a survival period. The outcome of failure is often death or some negative experience. Failure means death or a very negative experience. It can be, though, that failing is also a good thing, for example, when it leads to a cure for a disease. The failure could also be referred to as an event or an outcome, depending on whether it results in death.

We use two methods of estimation (the Maximum Likelihood Method and the Moment Method). However, the parametric model (Lindley Model) was used for survival analysis data. Additionally, all the corresponding results have been provided, and a comparison between the methods has been conducted.

Using measures provided in the goodness-of-fit used to retrieve how well the model fits the leukemia patients' data, the selected four criteria to determine the best model fit are (log-likelihood, BIC, AIC and MSE).

2. The main objective

The primary objective of this research is to utilise the Lindley parametric survival model analysis to develop the most effective statistical model that explains the relationship between the response variable, survival time, and a set of explanatory variables related to patients and the types of treatments used to enhance leukemia patient survival in Erbil City. The model's coefficients are estimated using the maximum likelihood estimation and moments method. Additionally, identifying the diagnostic characteristic that has the most considerable influence on a leukemia patient's survival time using the Lindley parametric survival model. The results should reveal the reliability and accuracy of each technique in a medical setting, particularly from a survival analysis perspective, where data is often subject to censoring. Findings from this study could significantly impact survival statistics and inform clinical decisions, providing oncologists and researchers with valuable insights into selecting the most suitable estimation method for a given set of cancer data.

3. Literature review

In 2008, Ghitany, Atieh and Nadarajah treated the mathematical properties of the Lindley distribution. Among the properties explored are moments, failure rate function, mean residual life function, mean deviations, ratios, and maximum likelihood estimation and simulation. One application to waiting time data at a bank is provided.[1]

In 2015, Bhati, Malik, and Vaman introduced a new distribution class that obtained the Lindley distribution via integral transform, offering flexible hazard shapes. Distributional properties, parameter estimation using the MLE method, and real data applications demonstrated a superior fit and usefulness in stress–strength reliability modelling.[2]

In 2016, Kartsonaki studied survival analysis techniques, censoring, Kaplan-Meier curves, log-rank tests, and Cox models. It was a clinically based article that provided a lucid overview of several statistical analyses employed in medical or diagnostic histopathology research.[3]

In 2019, Mawlood employed two advanced statistical methods to study the most significant factors affecting leukemia in Erbil city. The logistic regression and Cox regression. The results indicated that, despite the different regression coefficients in some cases of logistic regression and Cox regression, they have not identified the same variables that have an impact on the phenomenon. Moreover, the results indicated that the surgery is the most critical factor affecting the survival of leukemia patients in both methods.[4]

In 2021 (Asmaa and Ibtisam), the Asymmetric Laplace Distribution (AL), which has a fundamental and vital role in the development of mathematics and statistics, and its properties were applied in the financial field. The quadratic error loss function was compared with the maximum likelihood method, and the results showed that the Bayesian estimator for the skewness and measure parameters under the quadratic error loss function was better than the maximum likelihood method. [5]

In 2023, Iyad and Arshad employed maximum likelihood estimators and genetic algorithm estimators to estimate the parameters of the Gamma-Lindley distribution. The results indicated that the data fitted the Gamma-Lindley distribution and that the genetic algorithm estimators performed better than the maximum likelihood estimators, as determined by the AIC and BIC goodness-of-fit criteria. [6]

In 2023, Manal studied the reliability of the Lindley distribution. Two crucial methods were used to estimate the survival function: the maximum likelihood method and the method of moments. To enhance the capabilities of these two methods, a genetic algorithm was employed. Both methods demonstrated that the survival function exhibited a decreasing trend with increasing survival times for all methods. [7]

4. Background Information

This chapter introduces survival data analysis, focusing on the basic principles of survival data analysis in the context of a significant disease, Leukemia , and the fundamental principles and methods for modelling survival time data in the presence of censored observations. Also, the Lindley distribution and Lindley parametric survival model were defined and discussed. Additionally, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were applied to select the best model among the estimation methods.

4.1 Leukemia

Leukemia is a malignant neoplasm characterised by the heterogeneity of leukocytes and originating from the body's blood-forming tissues. HL is caused by acquired mutations in hematopoietic stem or progenitor cells, resulting in the dysregulated growth and proliferation of abnormal white blood cells. These abnormal cells infiltrate the bone marrow, where they suppress normal hematopoiesis, as well as infiltrate other tissues, leading to a myriad of systemic symptoms. The clinical features of this condition are primarily anaemia, thrombocytopenia, neutropenia, bone pain, organomegaly (liver/spleen), and increased susceptibility to infections. Established risk factors include exposure to ionising radiation, benzene and certain chemotherapeutic drugs. Other environmental and behavioural risk factors, such as smoking, viral infections, and exposure to organic solvents, are still being studied. Leukemia biology and treatment ramifications remain an area where further research into targeted therapies and personalised treatment should continue to focus on improving outcomes further. [8]

4.2 Survival analysis

This has included survival analyses on data where the time until an event, such as death, diagnosis, or recurrence of a condition, has been utilised. The resulting numbers are referred to as survival times. But if the end isn't death, the output data are the time to the event data.

It is employed to analyse data on the time to the event of interest. The response variable represents the time until an event occurs, often referred to as failure time, survival time, or event time.

Survival analysis is a statistical method, or technique, of data analysis in which the variable of interest is the time to an event. The amount of time between the beginning of the term when registration occurs and the actual event can range from days to weeks, months, or years.[9]

4.3 Survival Function

In addition to analysing most phenomena based on available information and data, the Survival function plays a dominant and practical role in many sciences, as time and the probability that a living being survives for at least a given time (t) are of great importance. Thus, most scholars have resorted to studies on this area of research, analysing and assessing the survival function.

$$S(t) = \Pr(T > t) = 1 - F(t) = \int_t^{\infty} f_{(u)}\, d_{(u)} \qquad (1)$$

Where u is an exploratory variable or (covariate), S(t) is the survival function, and it is assumed that T is a continuous random variable with probability density function (P.d.f) f(t), and the Cumulative Distribution Function (C.D.F). $F(t) = p_r(T \leq t)$, giving the probability that the event has occurred by duration t, S(t) is always one at t = 0.

The probability of surviving beyond time t, where $t \geq 0$ Thus, the survival function is a monotonically decreasing function.[3]

$$S(t) = p_r(T > t)\,, t \geq 0$$

$$S(t) = 1 - p_r(T > t)$$

$$S(t) = 1 - F(t) \qquad (2)$$

T: represents the survival time variable, or the time taken for the event to occur (time to events), which is the random variable, the death event that indicates the survival time until death occurs.

t: Survival time represents the specified time.

4.4 Hazard Function

The survival time T is given by the hazard function h (t), which is the conditional failure ratio; that is "the likelihood of failing within an infinitesimally small-time interval, given survival at the beginning of this interval", and the equation is:

$$\mathrm{h(t)} = \frac{\text{number of patients dying per unit time in the interval}}{\text{number of patients surviving at t}}$$

The hazard function is invaluable because it can describe not only the process of failure but also how the probability of the event varies over time.

The hazard function h(t) represents an individual's probability of death at time t after survival time. It represents the probability that an individual fails within a small interval (t, t+Δt) if they survive to the beginning of the interval. Which is written as:[10]

$$h_{(t)} = \lim_{\Delta t \to 0}[\frac{pr(t \le T \le t + \Delta t | T \ge t)}{\Delta t}]$$

$$h_{(t)} = \lim_{\Delta t \to 0}[\frac{pr(t \le T \le t + \Delta t) | pr(T \ge t)]}{\Delta t}$$

$$h_{(t)} = \lim_{\Delta t \to 0}[\frac{[F(t + \Delta t) - F(t)] | \Delta t)}{S(t)}]$$

$$h_{(t)} = \frac{\partial F(t) | \partial t}{s(t)}$$

$$h_{(t)} = \frac{f(t)}{s(t)} \quad (3)$$

4.5 Censoring Data

A further essential component of survival analysis is being able to participate in censoring, which refers to situations in which some information on the survival time of a patient is known but the exact survival time is unknown. Censorship may be left or right. Right censoring is the most prevalent form of censorship, but it is also the most prevalent type of "right" censoring. Censored observations create the unusual problem in survival analysis of not being able to "easily predict survival and hazard functions".[11]

4.6 Parametric distribution models for time to event

Survival analysis is concerned with particular dates, events, or lifetimes. The use of parametric survival models, such as the Weibull distribution, lognormal distribution, and log-logistic distribution, along with related inferential procedures, to describe lifetime distributions and their relationship with covariates. The benefit of this approach is that the estimates are easier to compute, and the resulting survival curves are smoother because they utilise information from the completely observed data. The disadvantage of parametric approaches is that they require other assumptions that might not be adequate. Parametric models involved proportional hazard measures and accelerated failure time (AFT) measures. In AFT models, the covariates are allowed to predict survival.[10]

4.6.1 Accelerated Failure Time

The AFT model is primarily employed in industrial applications and is scarce in survival data. The AFT model, if a suitable parametric form is chosen, provides a statistical model for survival data that is probabilistic and that is based on the survival curve rather than the hazard function. This model is called the accelerated failure time model because the word 'failure' refers to events such as death, illness, etc., and the word 'accelerated' refers to the factor that hastens these failures. This is referred to as the "accelerating factor". The AFT model is also referred to as the Logarithmic location scale model by Lawless (1982).

For $i = 1, \dots, n$ let T_i be the failure time for the ith individual and let X_i be the associated p-vector of covariates. The AFT model denotes that

$$\mathrm{Log}\, T_i = B_0 + B_i X_i + \varepsilon_i \quad (4)$$

Where B_0 is a p-vector of unspecified regression parameters and ε_i $(i = 1, \dots, n)$ are independent error terms with the common, but completely undefined, distribution.[10]

4.6.1.1 The Assumption of the AFT Model

The assumptions of AFT are thus as follows:

1. The distributional family is correctly specified.
2. In the absence of nonlinear and interaction terms, each $X_j$affects log (T).
3. Implicit in these assumptions is that δ is a constant independent of X.

An accelerated test environment is typically created by increasing the level of one or more stress variables, such as temperature or voltage.[10]

4.7 One-parameter Lindley Distribution

The Lindley distribution is a continuous probability distribution that was introduced by D.V. Lindley in 1958. Due to its convenience and flexibility, it has been extensively used in reliability analysis, survival analysis, actuarial science and queueing theory. The one-parameter Lindley distribution represents a mixture of an exponential and a gamma distribution. It has only one parameter but is more flexible than the exponential distribution, as it can accommodate a non-constant hazard rate; therefore, it is appropriate for lifetime data.[12]

For a positive real number $x > 0$, and shape/rate parameter $\theta > 0$, the PDF of the Lindley distribution can be defined as follows:

$$f(x;\theta) = \frac{\theta^2}{\theta+1}(1+x)e^{-\theta x}, \qquad x > 0 \tag{5}$$

- The constant $\frac{\theta^2}{\theta+1}$ ensures that the area under the curve integrates to 1 (i.e, it a proper probability distribution)
- The term $(1+x)$ also provides more flexibility than the standard exponential distribution.
- The exponential term $e^{-\theta x}$ represents the typical decay behaviour in lifetime models.[1]

The one-parameter Lindley distribution has a closed-form CDF that describes the cumulative probability of a positive continuous random variable. For a given shape parameter $\theta > 0$, the CDF is defined by the function:

$$F(x;\theta) = 1 - \left(1 + \frac{\theta x}{\theta+1}\right)e^{-\theta x} \tag{6}$$

valid for all $x > 0$. The function complies with the basic properties of a cumulative distribution: it is strictly increasing, continuous and differentiable in its domain.[13]

The mean (Average) of a random variable $X \sim Lindley(\theta)$, where $\theta > 0$, is another important measure of central tendency, representing the mean time to the occurrence of an event (such as failure or death). The closed form gives it:

$$E(X) = \frac{2+\theta}{\theta(\theta+1)} \tag{7}$$

The mean is in the exact dimensions as the variable of interest and also serves as a focal point in a reliability analysis, representing MTTF. It can be employed in moments estimation, model selection, and to compare the performance of competing lifetime distributions.[14]

The variance of the random variable $X \sim Lindley(\theta)$, where $\theta > 0$ measures the spread, or variability, of the distribution. It is derived from the second central moment and has a closed-form expression of:

$$Var(X) = \frac{\theta^3 + 4\theta^2 + 6\theta + 2}{\theta^2(\theta+1)^2} \tag{8}$$

Simply put, this formula applies to all positive θ, providing a finite and positive variance. As θ increases, the variance decreases; higher values of the shape parameter indicate a more peaked/less variable distribution. On the other hand, lower values of θ produce higher variances, thus implying more heterogeneous data.[12]

4.7.1 Survival Function Lindley Distribution

The Survival Function for a continuous random variable $X \sim Lindley(\theta)$ with $X > 0$, gives the estimated chance that the subject of interest will not experience the event by time x, and is a key

function in survival analysis and reliability theory. Mathematically, it is the complement of the cumulative distribution function, or CDF, defined as follows:

$$S(x;\theta) = \left(1 + \frac{\theta x}{\theta + 1}\right) e^{-\theta x}, \qquad x > 0 \tag{9}$$

The Lindley survival function is monotone decreasing and satisfies $S(0;\theta) = 1$, and converges to zero $x \to \infty$, This verifies that the chances of surviving in the long term eventually go to zero.

The survival function, therefore, provides a transparent and manageable representation of the survival or longevity of a given component, individual, or policy beyond a certain point, and thus has applications in various fields, including medical prognosis, engineering maintenance, and insurance risk assessment.[2]

4.7.2 Hazard Function Lindley Distribution

The underlying hazard function or failure rate function of the one-parameter Lindley distribution measures the instantaneous risk of failure at time. $x$given that the event has not yet occurred before x. The hazard function for a random variable $X \sim Lindley(\theta)$, where $\theta > 0$ can be expressed as the PDF divided by the survival function:

$$h(x;\theta) = \frac{f(x;\theta)}{S(x;\theta)} = \frac{\frac{\theta^2}{\theta+1}(1+x)e^{-\theta x}}{(1+\frac{\theta x}{\theta+1})e^{-\theta x}} = \frac{\theta^2(1+x)}{\theta+1+\theta x} \quad , \qquad x > 0 \tag{10}$$

This function expresses the instantaneous probability of an event occurring (such as failures, deaths, breakdowns, etc.) at a given time. x given that survival has occurred up to x. This expression is always positive and well-defined for > 0 , and it is clear from its form that this is an increasing function of x, which indicates an "ageing" process in that the longer a subject survives, the more likely they are to fail at the next moment. The Lindley hazard function is therefore an excellent candidate for modelling biological lifetimes, mechanical failure/wear-out, and non-constant risk systems.[15]

4.7.3 The Assumptions of the One-Parameter Lindley Distribution

1. Positive Data Values: All of the observed values, such as survival time, are positive and continuous. The Lindley distribution is only defined for $t > 0$.
2. Independence and Identical Distribution (i.i.d.)**:** The observations are assumed to be independent of each other and identically distributed.
3. Single Shape Parameter: The model is dependent on one positive parameter $\theta > 0$, which is used to control both the scale and shape of the distribution..
4. Monotonic Hazard Rate: The hazard rate of the Lindley distribution is increasing in general; hence, the risk of failure or death rises with time. [15]

4.8 Maximum Likelihood Estimation

The maximum likelihood method is fundamental to the analysis of accelerated test data. They are commonly used with many types of data and models. These approaches yield estimates and confidence intervals for model parameters, as well as other intriguing results.[16]

If $x_1, x_2, \dots, x_n$ are iid observations from a distribution that depends on the unknown parameters $\theta_1, \theta_2, \dots, \theta_m$ the likelihood function is given by:

$$L(\theta|x) = Pr(X_1 = x_1, x_2, \dots, X_n = x_n)$$
$$= f(x|\theta) = f(x_1|\theta) \times \dots \times f(x_n|\theta)$$

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta) \tag{11}$$

The application of MLE regarding the Lindley distribution is based on maximising the likelihood function, which for a given sample is formed using the probability density function of the Lindley

distribution. The likelihood function based on a random sample $(x_1, x_2, \dots, x_n)$ The probability density function from the Lindley distribution is the product of the individual densities. This is mathematically done by maximising the log-likelihood function concerning the parameter (θ), which is the scale parameter of the Lindley distribution; therefore, the estimation equation is obtained by differentiation and setting to zero.

$$\log L(\theta) = n \log\left(\frac{\theta^2}{1+\theta}\right) + \sum_{i=1}^{n} \log(1 + x_i) - \theta \sum_{i=1}^{n} x_i$$

This simplifies to:

$$\log L(\theta) = n(2 \log \theta - \log(1+\theta)) + \sum_{i=1}^{n} \log(1 + x_i) - \theta \sum_{i=1}^{n} x_i \quad (12)$$

Finding the maximum likelihood estimators for the Lindley distribution means solving the first order conditions in order to obtain analytically the values of the parameters, but the log-likelihood function is usually not solvable and it must therefore be approached through numerical methods like Newton-Raphson or the Expectation-Maximisation algorithm. Starting from a set of initial parameter estimates based on a priori knowledge or descriptive statistics, these algorithms iteratively refine estimates of θ until they converge. That MLE depends on relatively large samples and that it is susceptible to the initial values of the parameters are significant disadvantages in small-sample situations or when the likelihood surface is highly complicated and/or multimodal.[1]

4.9 Moment Estimation

MOM is a more straightforward approach, as it assumes that sample moments should equal their corresponding theoretical moments, providing a concise and computationally inexpensive method for parameter estimation. Moments of the Lindley distribution provide parameter estimates, as the sample moments must be equal to the theoretical moments, thereby offering a simpler alternative to Maximum Likelihood Estimation, which requires more complex computations.[17]

When applying MOM to the Lindley distribution, we must first establish the moments of the distribution. The first theoretical mean of the Lindley distribution, characterised by its parameter θ, plays a vital role in the MOM estimation procedure. This method involves setting this theoretical moment equal to the observed sample mean and deriving the estimation equation from there. In mathematical terms, if $(X1, X2, \dots, Xn)$ is a random sample from the Lindley distribution, then the moment is the mean of the sample, which is given by $(X) = X$. Solving for the value of θ when $\bar{X}$ is equal to the theoretical mean produces the MOM estimator for theta.

The MOM estimation method is a statistical tool among many, handy when computing facilities are restricted or as an initial assessment of parameter values.[18]

The first theoretical moment, or expected value of X, is given by:

$$E[X] = \mu(\theta) = \frac{2+\theta}{\theta(1+\theta)} \quad (13)$$

Derivation of the MoM Estimator:

Let $X_1, X_2, \dots, X_n$ be a random sample from the Lindley distribution. The sample mean is:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (14)$$

Equating the theoretical mean μ(θ) with the sample mean $\bar{X}$ gives:

$$\bar{X} = \frac{2+\theta}{\theta(1+\theta)}$$

$$\bar{X}\theta(1+\theta) = 2+\theta$$

$$\bar{X}\theta + \bar{X}\theta^2 = 2+\theta$$

$$\bar{X}\theta^2 + (\bar{X} - 1)\theta - 2 = 0$$

This is a quadratic equation in θ, which we solve using the quadratic formula:

$$\theta = \frac{-(\bar{X} - 1) \pm \sqrt{(\bar{X} - 1)^2 + 8\bar{X}}}{2\bar{X}}$$

Since $\theta > 0$, we select the positive root:

$$\hat{\theta}_{MoM} = \frac{1 - \bar{X} + \sqrt{(\bar{X} - 1)^2 + 8\bar{X}}}{2\bar{X}} \quad (15)$$

4.10 Measures of Model Selection

There are some criteria to identify the best model, given that several estimation methods provide accurate and good performance each on arbitrary data set:

4.10.1 Akaike's Information Criterion

The Akaike Information Criterion AIC is a useful tool for comparing the estimated algorithm's difficulty to how well it matches the data. (AIC) is a measure of the comparative accuracy of statistical models for a given collection of data, as well as an estimator of prediction error. AIC computes the quality of each model in contrast to the other models given a set of data models; the less information a model loses, the higher the level of its accuracy.

Akaike's Information Criterion is determined as follows:

$$AIC = -2loglikelihood + 2K \quad (16)$$

Where: K is the number of parameters in the model (the number of variables in the model plus the intercept). Log-likelihood is a measure of how well the model fits and is usually available from statistical output.[19]

4.10.2 The Bayesian Information Criterion

The Bayesian information criterion (BIC) is a model selection criterion for a finite set of models. Assume we want to find the model that best fits a given dataset. These models are usually not all of the same length. Schwarz established the Bayesian information criterion (BIC) in 1978, which serves as a model selection criterion: models with a lower BIC are selected.

Mathematically, BIC can be defined as:

$$BIC = -2loglikelihood + 2 * logN * k \quad (17)$$

Where L is the likelihood value, N is the total number of measurements, and k is the number of parameters being estimated. Model comparison using the BIC involves calculating the BIC for each of the models being compared. The model with the smallest BIC is considered to be the best model.[20]

4.10.3 The Mean Square Error

In statistical inference, the Mean Squared Error or MSE is a standard measure of an estimator's performance. It measures the mean-squared error of the estimates given the actual parameter value and combines information on bias and variance. The MSE for a particular estimator of a parameter$\hat{\theta}$,is given by:

$$\text{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \left(\text{Bias}(\hat{\theta})\right)^2 \quad (18)$$

For instance, in survival analysis and reliability studies, the one-parameter Lindley distribution is a valuable tool; thus, the MSE can significantly impact the comparison among various estimation techniques, including Maximum Likelihood Estimation, the method of Moments, and others.

The pdf of one-parameter Lindley distribution is:

$$f(t;\theta) = \frac{\theta^2}{\theta + 1}(1 + t)e^{-\theta t}$$

In a simulation study or using real data, MSE can be computed empirically as:

$$\widehat{MSE}(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta_i} - \theta)^2 \quad (19)$$

Where N is the total number of replications and $\hat{\theta_i}$is the estimate obtained in the $i-th$ replication.[21]

5. Results and Discussions:

In this part, the Lindley survival parametric model was utilised for survival analysis of leukemia patients' data, with two estimating methods: MLE and MOM. Additionally, all relevant results have been provided, along with a comparison of the two models. Statistical criteria (AIC, BIC, and MSE) were employed to determine the best model fit for our data. The data was analysed using the R program.

5.1 Data Collection

The data obtained for this leukemia study came from Erbil's Rizgari Hospital. 774 cases from all leukemia patients enrolled between September 1, 2021, and June 30, 2024, contributed to the data. The duration of survival is measured in months, starting from the patient's initial hospital admission and ending with their death or the date of their final hospital admission.

The reason for choosing the city of Erbil as the research site is due to the area's importance in cancer treatment and the high number of cancer patients, as well as related factors such as demographics, environment, and healthcare that may affect cancer patients in the city.

Additionally, the availability of clinical records and the collection of accurate information, with the assistance of the health institution, is necessary to obtain ethically sound information.

The study's prognostic factors, which were gathered for every patient, included the following covariates: (Age, Gender, Grade, Surgery, Chemotherapy, Radiotherapy, Hormone, Isotope treatment, Targeted treatment, Family history, Nationality, Occupation of the patient). Table (1) below lists the specific variables that were used and how they were categorised.

Table (1) variable categorisation

Variable	Describe Variable	Categorization	Count(N)	Rate	Status	
					No. of Death	No. of Alive
Age	Age of patients	2 - 29	11	1.42%	5	6
		3 - 37	81	10.47%	20	61
		38 - 45	196	25.32%	31	165
		46 - 53	182	23.51%	32	150
		54 - 61	163	21.06%	24	139
		62 - 69	93	12.02%	9	84
		70 - 77	34	4.39%	6	28
		78 - 125	14	1.81%	4	10
Gender	A patient which gender?	1=Male	207	26.74%	34	173
		2=Female	567	73.26%	97	470
Grade	Stage of Cancer	Stage 1	39	5.04%	4	35
		Stage 2	443	57.24%	78	365
		Stage 3	290	37.47%	49	241
		Stage 4	2	0.26%	0	2
Surgery	Have had surgery?	1=Yes	78	10.08%	9	69
		2=No	696	89.92%	122	574
Chemo	Do they use chemotherapy?	1=Yes	52	6.72%	2	50
		2=No	722	93.28%	129	593
Radio	Do they use radiotherapy?	1=Yes	84	10.85%	15	69
		2=No	690	89.15%	116	574
Hormone	Received hormone therapy?	1=Yes	149	19.25%	25	124
		2=No	625	80.75%	106	519
Isotope	Do they use isotope treatment?	1=Yes	738	95.35%	126	612
		2=No	36	4.65%	5	31
Target		1=Yes	565	73.00%	94	471

	Do they use targeted therapy?	2=No	209	27.00%	37	172
Family History	Cancer in patients family history	1=No	504	65.12%	79	425
		2=Yes	270	34.88%	52	218
Nationality	Citizenship	1=Iraqi	733	94.70%	124	609
		2=Arab(Not Iraqi)	38	4.91%	6	32
		3=Other	3	0.39%	1	2
Occupation	What kind of career does have?	1=Unemployed	460	59.43%	80	380
		2=Worker	26	3.36%	3	23
		3=Farmer	15	1.94%	2	13
		4=Employee	134	17.31%	22	112
		5=Professional	25	3.23%	5	20
		6=Child	3	0.39%	0	3
		7=Retirement	37	4.78%	5	32
		8=Other	74	9.56%	14	60
Status	The patient is alive or dead?	0=Event	131	16.93%		
		1=Censored	643	83.07%		
Time	Time diagnosed with cancer? (by month)					

Table (1) reveals that the age of diagnosis ranged from 2 to 125 years; the majority of patients (25.32%) were in the age group of 28 to 35, including 196 cases; out of them, 31 patients died, and 165 cases remained alive. A total of 774 patients (26.74%) were male, while 567 patients (73.26%) were female. Only two patients in the study proceeded to the fourth stage of leukemia cancer, and neither died from it. In the study, a total of 774 cases were analysed, with 78 undergoing surgery, 52 receiving chemotherapy, 84 undergoing radiotherapy, 149 undergoing hormone therapy, 36 undergoing isotope treatment, and 209 receiving targeted therapy. There were 270 patients (34.88%) with a family history of cancer, 52 of whom died. In terms of occupation, the majority of patients, 460 (59.43%), are unemployed; out of 774 patients in the research, 634 (66.9%) are still living, while 131 have died.

5.2 Distribution Fitting for the Data

Fitting the data to the studied distribution for survival time in our data is the first step in the process. To know that the dependent variable (survival Time of leukemia patients) follows the one-parameter Lindell distribution, we have applied the Goodness of Fit test by applying the $\chi 2$ test according to the following statistical hypothesis:

H0: The data fits one Parameters Lindley Distribution

H1: The data not fits one Parameters Lindley Distribution

Table (2) shows the results of the goodness-of-fit test of the data with the distribution used according to the following table:

Table (2) Goodness of Fit Test for Lindley Distributions

Distribution	chi-square	p Value	Decision
One Parameters Lindley Distribution	3.668	0.071	Not reject H0

Table (2) shows that the estimated value of chi-square is (3.668) and the p-value is (0.071), which is greater than the significance level (0.05). This indicates the null hypothesis cannot be rejected, and the survival time follows the one-parameter Lindley distribution.

5.2.1 Fitting the Lindley Survival Model using the Maximum Likelihood Estimation Method

The Lindley survival model was employed to demonstrate the impact of prognostic factors on survival time within the context of the data analysis. The regression coefficient in the predictive model, which operates 12 variables, indicates both the strength and direction of the relationship between the predictor factor and the response variable. The sign of a regression coefficient reveals whether the independent and dependent variables are positively or negatively related. The results of the Lindley Survival Model, employing the MLE method, are presented in Tables (3).

Table (3) Analysis of Fitting the Lindley Survival Model using the MLE Method

Parameters	β	Standard Error	95% Confidence Limits		P-Value	Z-Value
			Lower	Upper		
Constant	1.2414	0.8299	-0.3853	2.8680	0.1347	1.4957
Age	0.0272	0.0225	-0.0170	0.0714	0.2280	1.2056
Gender	0.1746	0.1039	-0.0290	0.3782	0.0927	1.6811
Grade	0.0434	0.0634	-0.0809	0.1678	0.4935	0.6847
Surgery	0.0300	0.1224	-0.2099	0.2700	0.8062	0.2453
Chemo	-1.8468	0.3515	-2.5358	-1.1579	0.0000	-5.2541
Radio	0.2247	0.1249	-0.0200	0.4694	0.0719	1.8000
Hormone	-0.3375	0.0927	-0.5192	-0.1558	0.0003	-3.6409
Isotope	0.4841	0.1956	0.1008	0.8674	0.0133	2.4755
Target	-0.2799	0.0830	-0.4427	-0.1172	0.0007	-3.3720
Family History	-0.2076	0.0760	-0.3566	-0.0586	0.0063	-2.7310
Nationality	0.3904	0.1278	0.1400	0.6408	0.0022	3.0556
Occupation	-0.0211	0.0195	-0.0593	0.0172	0.2802	-1.0799

The survival model can be written as follows:

$$\text{Log } T_i = B_0 + B_1X_1 + \cdots + B_iX_i + \varepsilon_i$$

We fit the survival model above to the data in Leukemia patients, the model will be:

$$\begin{aligned} LogT_i = {} & 1.2414 + 0.0272Age + 0.1746Gender + 0.0434Grade + 0.0300Surgery \\ & - 1.8468Chemo + 0.2247Radio - 0.3375Hormone \\ & + 0.4841Isotope - 0.2799Target - 0.2076FamilyHistory \\ & + 0.3904Nationality - 0.0211Occupation \end{aligned}$$

If we examine Table 4.3 above, the maximum likelihood coefficients for several parameters are statistically significant at the 5% level, as indicated; statistically significant coefficients indicate whether the relevant variable has a positive or negative effect on patients' survival times. We observe that each of the covariates (age, gender, grade, surgery, radio, isotope, and nationality) has a positive coefficient, while the other covariates are all negative. Based on the standard error, we can determine which beta has the best precision; the lowest value of standard error indicates the most accurate estimate.

The significant values (P-values) for the variables (Chemo=0.000, Hormone=0.0003, Isotope=0.0133, Target=0.0007, Family History=0.0063, and Nationality=0.0022) are all less than 0.05, which means that these variables are essential to the model we have chosen and are working on. And other variables are non-significant in the model because their P-values are greater than (0.05),

5.2.2 Fitting the Lindley Survival Model using the Moments Method

Using the Moments approach, the Lindley survival model has been applied to illustrate how prognostic factors affected survival time; the results are displayed in Table 4. According to the results, the covariates' (Chemo, Hormone, Target and Family History) computed coefficients are significant while their (p-values) are lower than the 5% levels. Meanwhile, the (Age, Gender, Grade, Surgery, Radio, Isotope, Nationality, Occupation) of the calculated coefficient are not significant, considering their p-values are significantly greater than the 5% value.

Table (4) Analysis of Fitting the Lindley Survival Model using MOM Method

Parameters	β	Standard Error	95% Confidence Limits		P-Value	Z-Value
			Lower	Upper		
Constant	-1.2617	2.2389	-5.6499	3.1266	0.5731	-0.5635
Age	0.0335	0.0672	-0.0981	0.1651	0.6176	0.4993
Gender	-0.1721	0.3462	-0.8506	0.5064	0.6191	-0.4971
Grade	-0.1357	0.1932	-0.5143	0.2430	0.4826	-0.7022
Surgery	-0.0159	0.4087	-0.8169	0.7851	0.9690	-0.0389
Chemo	1.9115	0.9020	0.1435	3.6795	0.0341	2.1191
Radio	0.1188	0.3494	-0.5661	0.8037	0.7338	0.3400
Hormone	0.5842	0.2643	0.0662	1.1022	0.0271	2.2104
Isotope	-0.7234	0.5432	-1.7880	0.3413	0.1830	-1.3317
Target	0.5217	0.2188	0.0928	0.9505	0.0171	2.3842
Family History	0.5965	0.2146	0.1758	1.0171	0.0055	2.7791

Nationality	-0.6299	0.3949	-1.4038	0.1441	0.1107	-1.5952
Occupation	0.0678	0.0622	-0.0541	0.1898	0.2757	1.0900

The survival model is as follows:

$$\text{Log } T_i = B_0 + B_1X_1 + \cdots + B_iX_i + \varepsilon_i$$

We fit the survival model above to the data in Leukemia patients.

$$\begin{aligned} LogT_i = &-1.2617 + 0.0335Age - 0.1721Gender - 0.1357Grade \\ &- 0.0159Surgery + 1.9115Chemo + 0.1188Radio \\ &+ 0.5842Hormone - 0.7234Isotope + 0.5217Target \\ &+ 0.5965FamilyHistory - 0.6299Nationality \\ &+ 0.0678Occupation \end{aligned}$$

5.3 Selection the best-fit model

The model with the lowest AIC, BIC, and MSE values is considered the optimal or best model. The result in Table (5) shows that the Lindley survival model with estimated coefficients by MLE has the lowest AIC, BIC and MSE values, hence a MLE method with 12 estimates parameters of Lindley survival model with the smallest value of (AIC=2381.843), (BIC= 2442.313) and (MSE=0.5281), under all performance measures tends to provide better estimates than MOM. MLE provides a superior fit, which is optimal and yields more efficient estimates of the parameters, leading to a more accurate and parsimonious model.

Table (5) Comparing Methods

Method	Number of parameters	Log Likelihood	AIC	BIC	MSE
MLE	12	-1177.921	2381.843	2442.313	0.5281
MOM	12	-1321.074	2668.147	2705.525	0.8306

6. Conclusions

Based on the results presented in the application part of this study, the main conclusions can be summarised as follows:

1. After checking the assumption of a one-parameter Lindley distribution for the response variable, survival time of leukemia cancer patients using the Pearson Chi-square test, the study concluded that the data satisfied the assumptions of this analysis.
2. The results of the likelihood ratio tests and the Goodness of fit tests indicated that the Lindley survival parametric model has a good fit to the dataset (the model adequately fits (describes) the data). In other words, the use of a Lindley survival parametric model to represent the effect of the studied explanatory variables on the survival time of leukemia cancer patients was successful and efficient.
3. The results of the maximum likelihood method estimation for leukemia patients showed that there are six explanatory variables (covariates) that have a significant effect on the survival time of leukemia cancer patients. These variables are: (Chemo, Hormone, Isotope, Target, Family History, and Nationality), but the effects and contributions of each variable are not the same.
4. The results of the study indicated that, through using the moments method to estimate parameters of the Lindley parametric survival model, four prognostic factors influence leukemia patients' survival, which are (Chemo, Hormone, Target, Family History)
5. The performance of the maximum likelihood and moments method in estimating the parameters of the Lindley survival parametric model in analysing the leukemia patient's data in Erbil city was evaluated using (AIC, BIC and MSE). The survival model estimation coefficients based on the maximum likelihood method appear to be the most suitable model.

7. Supplementary material

(None).

8. Author's Contributions

Ibrahim A. Othman: Designed the research. Kurdistan Ibrahim Mawlood.: Writing and editing.

9. Funding

(None).

10. Data availability statement

I collected my dataset from the Ministry of Health at Rizgary Hospital's Cancer Diseases Centre, based on daily records of cancer patients.

11. Acknowledgements

The authors would like to thank the Ministry of Health and the General Directorate of Health in Erbil for providing medical data, and Rizgary Hospital Centre for Cancer Diseases for supplying the leukemia dataset.

12. Conflict of interest

The authors declare no conflict of interest.

References

[1] Ghitany, M. E., Atieh, B., & Nadarajah, S. (2008). Lindley distribution and its application. Mathematics and computers in simulation, 78(4), 493-506.DOI: https://doi.org/10.1016/j.matcom.2007.06.007

[2] Bhati, D., Malik, M. A., & Vaman, H. J. (2015). Lindley–exponential distribution: properties and applications. Metron, 73, 335-357.DOI: https://doi.org/10.1007/s40300-015-0060-9

[3] Kartsonaki, C. (2016). Survival analysis. Diagnostic Histopathology, 22(7), 263-270. DOI: https://doi.org/10.1016/j.mpdhp.2016.06.005

[4] Mawlood, K. I. (2019). Using logistic regression and cox regression models to studying the most prognostic factors for leukemia patients. QALAAI ZANIST JOURNAL, 4(3), 705-724. DOI: https://doi.org/10.25212/lfu.qzj.4.3.20

[5] Abdullah, I. K., & Rady, A. K. (2021). Comparison of methods for estimating the parameters of the asymmetric Laplace distribution using the quadratic loss function and the maximum possibility method. Journal of Administration and Economics, 46(130).DOI: https://doi.org/10.31272/jae.i130.32

[7 Hassan, A. H., & Shamal, I. H. (2023). Using genetic algorithm to estimate gamma Lindley distribution parameters. Journal of Administration and Economics, 48(141).DOI: https://doi.org/10.31272/jae.i141.1007

[8] Manal M.R. (2023).Using the genetic algorithm to improve the survival function estimates of the Frechet-Weibull exponential distribution mixed model with a practical application. DOI: https://doi.org/10.31272/jae.i141.1012

[9] Zhang, T. D., Chen, G. Q., Wang, Z. G., Wang, Z. Y., Chen, S. J., & Chen, Z. (2001). Arsenic trioxide, a therapeutic agent for APL. Oncogene, 20(49), 7146-7153. DOI: https://doi.org/10.1038/sj.onc.1204762

[10] Redha, S. M., & Hadia, A. T. A. (2020). Estimate the Survival Function By Using The Genetic Algorithm. Journal of Economics and Administrative Sciences, 26(122). DOI: https://doi.org/10.33095/jeas.v26i122.2018

[11] Austin, P. C. (2017). A tutorial on multilevel survival analysis: methods, models and applications. International Statistical Review, 85(2), 185-203.DOI: https://doi.org/10.1111/insr.12214

[12] Harrell, F. E. (2001). Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis (Vol. 608). New York: springer.DOI: https://doi.org/10.1007/978-1-4757-3462-1

[13] Lindley, D. V. (1970). The estimation of many parameters. ETS Research Bulletin Series, 1970(1), i-20. DOI: https://doi.org/10.1002/j.2333-8504.1970.tb00411.x

[14] Shanker, R., Sharma, S., & Shanker, R. (2013). A two-parameter Lindley distribution for modeling waiting and survival times data. Applied Mathematics, 4(2), 363-368. DOI: http://dx.doi.org/10.4236/am.2013.42056

[15] Ramos, P. L., & Louzada, F. (2016). The generalized weighted Lindley distribution: Properties, estimation, and applications. Cogent Mathematics, 3(1), 1256022. DOI: https://doi.org/10.1080/23311835.2016.1256022

[16] Bhati, D., Sastry, D. V. S., & Qadri, P. M. (2015). A new generalized Poisson-Lindley distribution: Applications and properties. Austrian Journal of Statistics, 44(4), 35-51.DOI: https://doi.org/10.17713/ajs.v44i4.54

[17] Van Den Hout, A. (2016). Multi-state survival models for interval-censored data. Chapman and Hall/CRC. DOI: https://doi.org/10.1201/9781315374321

[18] Khan, M. J. S., Sharma, A., & Iqrar, S. (2020). On moments of Lindley distribution based on generalized order statistics. American Journal of Mathematical and Management Sciences, 39(3), 214-233. DOI: https://doi.org/10.1080/01966324.2020.1718568

[19] Sultan, K. S., & Al-Thubyani, W. S. (2016). Higher order moments of order statistics from the Lindley distribution and associated inference. Journal of Statistical computation and Simulation, 86(17), 3432-3445. DOI: https://doi.org/10.1080/00949655.2016.1163361

[20] Moore, D. F. (2016). Applied survival analysis using R (Vol. 473, pp. 1-10). Cham: Springer. DOI: https://doi.org/10.1007/978-3-319-31245-3
[21] Ibrahim, J. G., Chen, M. H., & Sinha, D. (2013). Bayesian survival analysis. Springer Science & Business Media. DOI: https://doi.org/10.1201/b16248
[22] Balan, T. A., & Putter, H. (2020). A tutorial on frailty models. Statistical methods in medical research, 29(11), 3424-3454.DOI: https://doi.org/10.1177/0962280220921889

https://doi.org/10.31272/jae.i150.1447
https://admics.uomustansiriyah.edu.iq
P-ISSN: 1813–6729 E-ISSN: 2707-1359

تركيب نموذج ليندلي للبقاء باستخدام طريقة الاحتمال الأعظم وطريقة العزوم لتحليل بيانات البقاء على قيد الحياة لمرضى سرطان الدم في أربيل_العراق.

ابراهيم عبدالخالق عثمان
قسم الإحصاء، كلية الإدارة والاقتصاد، جامعة صلاح الدين، مدينة أربيل، كوردستان، العراق.
Email: ibrahim.othman@su.edu.krd, ORCID: https://orcid.org/0009-0000-2179-9024
كوردستان ابراهيم مولود
قسم الإحصاء، كلية الإدارة والاقتصاد، جامعة صلاح الدين، مدينة أربيل، كوردستان، العراق.
Email: kurdistan.mawlood@su.edu.krd, ORCID: https://orcid.org/0000-0002-1612-1996

المستخلص

استخدمت الدراسة توزيع ليندلي لتقدير نماذج البقاء البارامترية باستخدام بيانات بقاء سرطان الدم في مدينة أربيل بالعراق. درسنا طريقتي تقدير لنمذجة بيانات وقت الحدث: تقدير الاحتمال الأعظم (MLE)، وطريقة العزوم (MoM). ينتج طريقة MLE تقديرات فعالة بشكل مقارب، بينما باستخدام لحظات العينة، توفر MoM نهجًا أبسط وغير تكراري. استخدمت اختبارات جودة الملاءمة، ومعايير معلومات أكايكي (AIC)، ومعيار المعلومات البايزي (BIC)، ومتوسط خطأ التربيع (MSE) لتقييم أداء النموذج. أظهرت النتائج أن MLE يتفوق على MoM من حيث الدقة والمرونة، لاسيما مع البيانات الخاضعة للرقابة. ومع ذلك، أظهرت النتائج أهمية توزيع ليندلي في تحليل البقاء الطبي وتوفر نظرة ثاقبة لأنماط بقاء سرطان الدم في مدينة أربيل. تُضيف هذه الدراسة إلى تنمية الأدبيات حول نمذجة البقاء البارامترية وتساعد في اتخاذ القرارات السريرية لعلاجات سرطان الدم. تم الحصول على بيانات هذه الدراسة من مستشفى رزكاري في مدينة أربيل. وتم الحصول على النتائج باستخدام الحزوم الإحصائية (Mat-lab V. 23.2 و R Software V. 2025.05.1+513).

معلومات البحث

تواريخ البحث:
التقديم: 23 /06/ 2025
المراجعة: 01 / 11 / 2025
قبول النشر: 06 / 11 / 2025
نشر الكتروني: 01 / 12 / 2025
تسلسل الصفحات: 39 - 53

الكلمات المفتاحية:
تحليل البقاء، نموذج ليندلي البارامتري للبقاء، معيار معلومات أكايكي (AIC)، سرطان الدم.

المراسلة:
أسم الباحث:كوردستان ابراهيم مولود

Email:
kurdistan.mawlood@su.edu.krd